Maximum-Likelihood Estimation of the Parameters of a Multivariate Normal Distribution*

T. W. Anderson and I. Olkin
Department of Statistics
Stanford University
Stanford, California 94305-2195

Submitted by George P. H. Styan

ABSTRACT

This paper provides an exposition of alternative approaches for obtaining maximum-likelihood estimators (MLE) for the parameters of a multivariate normal distribution under different assumptions about the parameters. A central focus is on two general techniques, namely, matrix differentiation and matrix transformations. These are systematically applied to derive the MLE of the means under a rank constraint and of the covariances when there are missing observations. Derivations using induction and inequalities are also included to illustrate alternative methods. Other examples, such as a connection with an econometric model, are included. Although the paper is primarily expository, some of the proofs are new.

1. INTRODUCTION

The multivariate normal distribution has served as a central distribution in much of multivariate analysis. The statistical goal is to obtain maximum-likelihood estimators (MLE) for the means and covariances. Particular applications often impose a structure or constraints on the parameters that sometimes make the maximizations more difficult.

Alternative techniques have been developed to obtain maximum-likelihood estimators. It is important to note that no single method is a panacea that

*Research supported in part by the Office of Naval Research Contract N00014-75-C-0442 (NR-042-034), Army Research Office Contract DAAG 29-82-K-0156, and the National Science Foundation Grant MCS 78-07736. The first author was Mitchell Visiting Professor of Economics at Columbia University and sabbaticant at the IBM Systems Research Institute while this paper was being completed.

52 Vanderbilt Ave., New York, NY 10017 0024-3795/85/$3.30

readily provides answers for all models. Certain techniques are designed to handle very specific models, and some may be better suited for one problem than for another. In this exposition we illustrate two particular approaches, namely, (1) differentiation and (2) matrix transformations. This does not imply that these methods are the only ones that can be used or that they are optimal. However, they are basic techniques and do apply to a wide class of models. We include other methods, such as induction and inequalities, to show how alternatives can also be used.

Setting first-order derivatives equal to zero is, of course, a fundamental method for finding extremals. When constraints are present, the method of Lagrangian multipliers can be used. (If the constraints are complicated, we may have to resort to numerical solutions rather than closed-form expressions.) In order to deal with extremal problems involving matrix functions, a calculus of matrix differentiation has been developed over the years. Two early texts that discuss matrix differentiation are by Frazer, Duncan, and Collar [20, Chapter 2] and Bodewig [13, Chapter 1]. More recently many variants have appeared: Dwyer and Macphail [18], Deemer and Olkin [16], Dwyer [17], Neudecker [34], McDonald and Swaminathan [31], Tracy and Singh [56], Nel [33], Rogers [45], and Graham [24].

Because the multivariate normal density is a function of both the trace and the determinant of a positive definite matrix, it is sometimes advantageous to maximize sequentially over the parameters, rather than simultaneously. By so doing, one can choose the second-stage maximization to be either over the trace or over the determinant, whichever provides the greater simplification.

Matrix inverses may not exist in models involving rectangular or singular matrices. In such instances generalized inverses can be used. Computer programs for generalized inverses are now available, so that the computational problems are no longer serious. For references on generalized inverses see [14], [44], or [12].

Models that involve patterns or relationships among the parameters can, at times, be resolved by special methods. Anderson [4, 5] provides an iterative procedure for obtaining the MLE of the covariances when the covariance matrix or its inverse is a linear function of other parameters. Powerful general methods for dealing with covariances that have a particular algebraic structure have been developed by Andersson [6] and Erlandsen [19]. This structure is exemplified by compound symmetry [57] and circular symmetry [38, 37]. Other papers in the area of patterned matrices include those, for example, of Arnold [7, 8] and Szatrowski [50, 51].

Because the normal distribution can be parametrized to be a member of the exponential family, the general theory of exponential families can be applied in special cases. For a discussion of the exponential family and the specialization to the normal distribution see [9, Chapter 9].

A key feature in the use of matrix transformations is that it provides a mechanism for reducing a model to a simpler canonical form. This aspect is emphasized throughout this paper.

NOTATION. A matrix A with m rows and n columns is denoted by $A: m \times n$; vectors $a = (a_1, \ldots, a_n)$ denote row vectors. D_a or $D(a_1, \ldots, a_n)$ denotes a diagonal matrix with diagonal elements $a_1, \ldots, a_n$. The determinant and trace are denoted $|A|$ and $\operatorname{tr} A$, respectively. For a $(k+l)\times(k+l)$ matrix A partitioned $A = (A_{ij})$, $i, j = 1, 2$, $A_{11}: k \times k$, $A_{22}: l \times l$, the Schur complements $A_{ii} - A_{ij}A_{jj}^{-1}A_{ji}$ are denoted $A_{ii \cdot j}$ whenever the inverses exist.

Positive and nonnegative definiteness of A are denoted as $A > 0$ and $A \geqslant 0$, respectively. The real characteristic roots $\lambda_1, \ldots, \lambda_n$ of a symmetric matrix are ordered $\lambda_1 \geqslant \cdots \geqslant \lambda_n$.

Greek letters generally denote parameters; Latin letters refer to sample values. The MLE of a parameter θ is often denoted by $\hat{\theta}$.

2. ESTIMATING THE COVARIANCES OF A MULTIVARIATE NORMAL DISTRIBUTION

One of the most basic problems in multivariate analysis is that of finding maximum-likelihood estimators of the mean μ and covariance matrix Σ of a normal p-variate distribution based on N p-dimensional vector observations $x_1, \ldots, x_N$. We assume that $N > p$ so that the sample covariance matrix is positive definite with probability 1 (given that Σ is positive definite). By sufficiency we can confine ourselves to a consideration of the joint distribution of the mean $\bar{x}$ and the sample covariance matrix

$$V = \frac{1}{N}\sum_{1}^{N}(x_\alpha - \bar{x})'(x_\alpha - \bar{x}). \tag{2.1}$$

Since $\bar{x}$ and V are independently distributed, and $\bar{x}$ has a normal distribution with mean μ and covariance Σ/N, it is straightforward to show that $\bar{x}$ is the MLE of μ. The logarithm of the concentrated likelihood is $-\frac{1}{2}pN\log 2\pi + \frac{1}{2}NK$, where K, the kernel of the concentrated likelihood, can be written as

$$f(\Sigma; V) = -\log|\Sigma| - \operatorname{tr}\Sigma^{-1}V. \tag{2.2}$$

The problem is to maximize $f(\Sigma; V)$ with respect to positive definite matrices Σ. Because $\Psi = \Sigma^{-1}$ is a one-to-one transformation of Σ, an

equivalent problem is to maximize

$$g(\Psi; V) = \log|\Psi| - \operatorname{tr} \Psi V \tag{2.3}$$

with respect to positive definite Ψ. (This is Lemma 3.2.3 of [3].)

REMARK. It is somewhat surprising that early reference to the fact that V is the MLE of Σ is elusive. But the general result is implicit in the work of Wilks [59, p. 476] dealing with likelihood-ratio tests.

One of the most common methods for finding the MLE of the covariance matrix is based on differentiation. We carry out this derivation (Section 2.1) in two ways: differentiation with respect to the elements of Σ and with respect to the elements of Σ^{-1}. These involve somewhat different arguments.

A second general technique frequently used is that of matrix transformations. Several different transformations can be used, and in Section 2.2 we discuss each of the alternatives.

It is natural in seeking an extremum to try to bound the likelihood, and then show that the bound is achieved. Such a method is the essence of Section 2.3.

Very often an inductive proof can be used to advantage—in particular, in that it may avoid some of the analytic complexities. This method is exhibited in Section 2.4.

The multivariate normal distributions constitute an exponential family of distributions and can be given a canonical parametrization. Barndorff-Nielsen [9, Chapter 9] shows that the likelihood function is log concave in that parametrization and has a unique maximum.

2.1. *The Method of Differentiation*

To obtain the derivative equations for (2.2) and (2.3) we use differential forms. (See the references on matrix differentiation in the Introduction.)

The needed facts are:

FACT 1.

$$(d\Sigma^{-1}) = -\Sigma^{-1}(d\Sigma)\Sigma^{-1}.$$

FACT 2.

$$(d|\Sigma|)_{i,j} = (2 - \delta_{ij})\Sigma_{ij}(d\sigma_{ij}).$$

Here Σ_{ij} is the cofactor of σ_{ij}, and δ_{ij} is Kronecker's delta. (These equations can be thought of as devices to keep account of the partial derivatives.)

2.1.1. Differentiation with Respect to the Elements of Σ. The function $f(\Sigma; V)$ is neither convex nor concave in Σ. However, $f(\Sigma; V) \to -\infty$ as Σ approaches the boundary of positive definite matrices, that is, as the smallest characteristic root of Σ approaches zero or as one or more elements increases without bound. Therefore, a maximum exists in the set of positive definite matrices. The derivative equations (obtained below) have only one solution; consequently the maximum is unique.

Differentiation of $f(\Sigma; V)$ defined by (2.2) yields

$$d[f(\Sigma; V)]_{ij} = (2 - \delta_{ij})\left\{ -\frac{\Sigma_{ij}}{|\Sigma|} d\sigma_{ij} + (\operatorname{tr} \Sigma^{-1} E_{ij} \Sigma^{-1} V)\, d\sigma_{ij} \right\} = 0, \qquad i \leqslant j, \quad (2.4)$$

where E_{ij} is a matrix with 1 in the (i, j)th position and 0 elsewhere, and Σ_{ij} is the cofactor of σ_{ij}. Equation (2.4) can be expressed as a matrix equation

$$-\Sigma^{-1} + \Sigma^{-1} V \Sigma^{-1} = 0,$$

which has the unique solution $\hat{\Sigma} = V$.

This approach is discussed by Smith [47] and used in the book by Kshirsagar [27].

2.1.2. Differentiation with Respect to the Elements of Σ^{-1}. That $g(\Psi; V)$ is strictly concave in Ψ follows from the linearity of $\operatorname{tr} \Psi V$ and the well-known result that $\log|\Psi|$ is concave (see [11, p. 128]). Since $g(\Psi; V) \to -\infty$ on the boundary of positive definite matrices, a maximum of $g(\Psi; V)$ with respect to Ψ exists and is unique.

Differentiating $g(\Psi; V)$ defined by (2.3) yields

$$d[g(\Psi; V)]_{ij} = (2 - \delta_{ij})\left\{ \frac{\Psi_{ij}}{|\Psi|} d\psi_{ij} - (\operatorname{tr} E_{ij} V)\, d\psi_{ij} \right\} = 0, \qquad i \leqslant j. \quad (2.5)$$

Equation (2.5) can be expressed quite simply as

$$\Psi^{-1} - V = 0,$$

which has the unique solution $\hat{\Psi} = V^{-1}$.

This approach is used in the books by Anderson [3] (1st ed.), Rao [43], and Mardia, Kent, and Bibby [29].

The contrast between the methods of Sections 2.1.1 and 2.1.2 shows that although it may be more natural to maximize with respect to the elements of Σ, a considerable simplification is achieved by maximizing with respect to the elements of Σ^{-1}.

2.2. *The Method of Matrix Transformations*

The functions $f(\Sigma; V)$ and $g(\Psi; V)$ can be written in a canonical form. For any factorization $V = CC'$, where C is a square nonsingular matrix, let

$$\tilde{\Sigma} = C^{-1}\Sigma C'^{-1}, \qquad \tilde{\Psi} = C'\Psi C. \tag{2.6}$$

Then (2.2) and (2.3) become

$$f(\Sigma; V) = -\log|V| - \log|\tilde{\Sigma}| - \operatorname{tr}\tilde{\Sigma}^{-1},$$

$$g(\Psi; V) = -\log|V| + \log|\tilde{\Psi}| - \operatorname{tr}\tilde{\Psi}.$$

Since C is known, maximization with respect to $\tilde{\Sigma}$ or $\tilde{\Psi}$ is equivalent to maximization with respect to Σ or Ψ, respectively. Further, the term $-\log|V|$ is a constant, so that we need only consider

$$f(\Sigma; I) = -\log|\Sigma| - \operatorname{tr}\Sigma^{-1}, \tag{2.7}$$

$$g(\Psi; I) = \log|\Psi| - \operatorname{tr}\Psi \tag{2.8}$$

as our starting points. (For simplicity of notation we omit the tildes.)

A variety of representations for a $p \times p$ symmetric matrix can be used. Three factorizations that are well known and often used are the following.

FACT 3. *If $H > 0$, then there exists a unique lower (or alternatively upper) triangular matrix $T = (t_{ij})$ such that $H = TT'$, where $t_{ii} > 0$, $i = 1, \ldots, p$.*

The elements t_{ij} of T are called rectangular coordinates in the statistical literature and were introduced from a geometric perspective by Mahalanobis, Bose, and Roy [28]. Rectangular coordinates were used earlier by Bartlett [10]. The factorization was given by Toeplitz [55] and in an equivalent form by Schmidt [46].

FACT 4. *If $H > 0$, then $H = DRD$, where $D = D\left(\sqrt{h_{11}}, \ldots, \sqrt{h_{pp}}\right)$, $h_{ii} > 0$, $i = 1, \ldots, p$, and R is a correlation matrix (that is, R is positive definite with unit diagonal elements).*

FACT 5. *If H is symmetric, then there exists an orthogonal matrix G such that $H = GD_dG'$, where $D_d = D(d_1, \ldots, d_p)$ and $d_1 \geq \cdots \geq d_p$ are the ordered characteristic roots of H. If $H > 0$, then $d_p > 0$.*

By using each of these transformations, we show that the original problem reduces to the univariate problem

$$\operatorname*{Max}_{z>0} (\log z - z), \tag{2.9}$$

which is readily solved, since $\log z - z$ is concave and has the unique maximum of -1 at $z = 1$.

2.2.1. Transformation to Rectangular Coordinates. After an application of the transformation of Fact 3 with $\Psi = TT'$, $T = (\tau_{ij})$, lower triangular, the maximization of $g(\Psi; I)$ becomes

$$\operatorname*{Max}_{\Omega} \left\{ \sum_i (\log \tau_{ii}^2 - \tau_{ii}^2) - \sum_{i>j} \tau_{ij}^2 \right\}, \tag{2.10}$$

where $\Omega = \{T: \tau_{ii} > 0, \ -\infty < \tau_{ij} < \infty, \ i > j; \ i, j = 1, \ldots, p\}$. Clearly, the maximum over τ_{ij}, $i > j$, occurs at $\tau_{ij} = 0$, so that (2.10) reduces to a sum of terms like (2.9).

2.2.2. Transformation by a Scaling Matrix. After an application of the transformation of Fact 4 with

$$\Psi = DPD, \qquad P = (\rho_{ij}), \qquad D = D\left(\sqrt{\psi}_{11}, \ldots, \sqrt{\psi}_{pp}\right),$$

the maximization of $g(\psi; I)$ becomes

$$\operatorname*{Max}_{\substack{\psi_{ii}>0 \\ P>0}} \left\{ \sum_i (\log \psi_{ii} - \psi_{ii}) + \log|P| \right\} = \operatorname*{Max}_{\psi_{ii}>0} \sum_i (\log \psi_{ii} - \psi_{ii}) + \operatorname*{Max}_{P>0} \log|P|. \tag{2.11}$$

We assert that $\log|P| \leqslant 0$, with equality if and only if $P = I$, so that (2.11) reduces to a sum of terms like (2.9).

To prove the assertion, we can use Hadamard's inequality $|P| \leqslant \prod_{i=1}^{p} \rho_{ii} = 1$, with equality for $P = I$. Alternatively, from Fact 3, let $P = UU'$, where $U = (u_{ij})$ is lower triangular with $\sum_{\alpha=1}^{i} u_{i\alpha}^2 = 1$, so that $|P| = \prod_{i=1}^{p} u_{ii}^2 \leqslant 1$.

2.2.3. Transformation to Characteristic Roots. After an application of the transformation of Fact 5 with $\Psi = \Gamma D_\delta \Gamma'$, $D_\delta = D(\delta_1, \ldots, \delta_p)$, the maximization of $g(\Psi; I)$ becomes

$$\operatorname*{Max}_{\delta_i > 0} \left\{ \sum_i (\log \delta_i - \delta_i) \right\},$$

which is of the form (2.9).

Use of Fact 5 in this context was suggested by Anderson [3, 1st ed., Problem 4, Chapter 3] and used by Watson [58]; it is the essence of the method of Khatri [26] and Tamhane [52]. (C. M. Theobald has pointed out to us that there is an inconsequential error in Tamhane's letter of treating $A\Sigma^{-1}$ as symmetric.) This transformation is used in the books by Giri [22] and Muirhead [32].

2.2.4. Simultaneous Reduction of Two Matrices. Another factorization that is frequently useful in reducing some models to a canonical form is the simultaneous diagonalization of two positive definite matrices.

FACT 6. *If V and Σ are $p \times p$ positive definite matrices, then there exists a nonsingular matrix L such that*

$$V = LL', \qquad \Sigma = LD_\lambda L', \tag{2.12}$$

where $D_\lambda = D(\lambda_1, \ldots, \lambda_p)$ and $\lambda_1 \geqslant \cdots \geqslant \lambda_p > 0$ are the ordered roots of $|\Sigma - \lambda V| = 0$.

An application of the transformation (2.12) to (2.2) yields

$$f(\Sigma; V) = -\log|V| - \sum_{i=1}^{p} \left(\log \lambda_i + \frac{1}{\lambda_i} \right),$$

which is maximized for $\hat{\lambda}_i = 1$, $i = 1, \ldots, p$. Consequently, $D_{\hat{\lambda}} = I$ and $\hat{\Sigma} = V$.

2.3. A Proof Based on an Inequality

A result of von Neumann [36] implies that

$$\operatorname{tr} \Psi V \geqslant \sum_{1}^{p} \eta_i \nu_{p-i+1}, \tag{2.13}$$

where $\eta_1 \geqslant \cdots \geqslant \eta_p$ and $\nu_1 \geqslant \cdots \geqslant \nu_p$ are the ordered characteristic roots of Ψ and V, respectively. Equality in (2.13) is achieved if and only if the characteristic vector of Ψ corresponding to η_i is equal to the characteristic vector of V corresponding to ν_{p-i+1}, $i = 1, \ldots, p$.

Applying (2.13) to (2.3) yields

$$g(\Psi; V) \leqslant \sum_{1}^{p} (\log \eta_i - \eta_i \nu_{p-i+1}),$$

with equality when the characteristic vectors of Ψ are identical to those of V. Then $(\log \eta_i - \eta_i \nu_{p-i+1})$ is maximized at $\eta_i = 1/\nu_{p-i+1}$, $i = 1, \ldots, p$, and $g(\Psi; V)$ is maximized at $\hat{\Psi} = V^{-1}$.

This method is used by Theobald [54].

2.4. An Inductive Proof

If $p = 1$, then $g(\Psi; I) = \log \Psi - \Psi$, which is maximized at $\Psi = 1$. For general p partition Ψ as

$$\Psi = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}, \qquad \Psi_{11} : (p-1) \times (p-1).$$

We wish to show that if $g(\Psi_{11}; I_{p-1})$ given by (2.8) is maximized at $\Psi_{11} = I_{p-1}$, then $g(\Psi; I_p)$ is maximized at $\Psi = I_p$.

A consequence of $\Psi > 0$ is that $\Psi_{11} > 0$ and $\Psi_{22 \cdot 1} > 0$, and we may write

$$|\Psi| = |\Psi_{11}| \, \Psi_{22 \cdot 1}.$$

Consequently, the maximization of $g(\Psi; I_p)$ becomes

$$\operatorname*{Max}_{\substack{\Psi_{11}, \Psi_{12}, \Psi_{22} \\ \Psi > 0}} \left\{ (\log |\Psi_{11}| - \operatorname{tr} \Psi_{11}) + \log(\Psi_{22} - \Psi_{21} \Psi_{11}^{-1} \Psi_{12}) - \Psi_{22} \right\}. \tag{2.14}$$

For fixed Ψ_{11} and Ψ_{22}, the maximum with respect to Ψ_{21} is achieved at $\Psi_{21}=0$, so that (2.14) reduces to

$$\max_{\substack{\Psi_{11}>0\\ \Psi_{22}>0}} \{(\log|\Psi_{11}|-\operatorname{tr}\Psi_{11})+(\log\Psi_{22}-\Psi_{22})\}$$

$$=\max_{\Psi_{11}>0}(\log|\Psi_{11}|-\operatorname{tr}\Psi_{11})+\max_{\Psi_{22}>0}(\log\Psi_{22}-\Psi_{22}).$$

The second maximization follows from (2.9) and occurs at $\Psi_{22}=1$; by the inductive hypothesis the first maximum occurs at $\Psi_{11}=I_{p-1}$.

3. ESTIMATING THE COVARIANCES FOR A MODEL WITH MISSING OR ADDITIONAL OBSERVATIONS

We show below that several statistical models lead to the canonical form

$$f(\Sigma)\equiv -N\log|\Sigma|-N\operatorname{tr}\Sigma^{-1}V-M\log|\Sigma_{11}|-M\operatorname{tr}\Sigma_{11}^{-1}W, \qquad (3.1)$$

where $\Sigma=(\Sigma_{ij})$, $i,j=1,2$, $\Sigma_{11}: k\times k$, $\Sigma_{22}:(p-k)\times(p-k)$, $V: p\times p$ is positive definite, $W: k\times k$ is positive semidefinite, and M and N are nonnegative constants. The function $f(\Sigma)$ is to be maximized over the region $\{\Sigma:\Sigma>0\}$.

If we let $\Psi=\Sigma^{-1}$ be partitioned conformably with Σ, and note that $\Sigma_{11}^{-1}=\Psi_{11\cdot 2}$, then (3.1) can be written as

$$g(\Psi)=N\log|\Psi|-N\operatorname{tr}\Psi V+M\log|\Psi_{11\cdot 2}|-M\operatorname{tr}\Psi_{11\cdot 2}W, \qquad (3.2)$$

which is to be maximized over the region $\{\Psi:\Psi>0\}$, or equivalently, over the region $\{\Psi_{11\cdot 2},\Psi_{12},\Psi_{22}:\Psi_{11\cdot 2}>0,\ \Psi_{22}>0\}$.

We now show how the canonical forms (3.1) and (3.2) arise.

Suppose that $(x,y)=(x_1,\ldots,x_k,y_1,\ldots,y_l)$ has a $(k+l)$-dimensional multivariate normal distribution where the mean vector of x is known, say $Ex=0$, but the mean $Ey=\mu$ of y is unknown. The covariance matrix Σ of (x,y) is partitioned as $\Sigma=(\Sigma_{ij})$, $i,j=1,2$, with $\Sigma_{11}: k\times k$ and $\Sigma_{22}: l\times l$. For a sample of size N, let $(\bar{x},\bar{y})$ denote the mean vector. (For simplicity let $\Psi=\Sigma^{-1}$ be partitioned conformably.) The loglikelihood function that in-

volves the means is proportional to

$$(\bar{x}, \bar{y}-\mu)\Sigma^{-1}(\bar{x}, \bar{y}-\mu)' = (\bar{x}, \bar{y}-\mu)\begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}\begin{pmatrix} \bar{x}' \\ \bar{y}'-\mu' \end{pmatrix}$$

$$= \left(\bar{y}-\mu+\bar{x}\Psi_{12}\Psi_{22}^{-1}\right)\Psi_{22}\left(\bar{y}-\mu+\bar{x}\Psi_{12}\Psi_{22}^{-1}\right)' + \bar{x}\Psi_{11\cdot 2}\bar{x}'. \tag{3.3}$$

For fixed Σ or Ψ, the minimizing μ is

$$\hat{\mu} = \bar{y} + \bar{x}\Psi_{12}\Psi_{22}^{-1}, \tag{3.4}$$

so that the minimum of (3.3) is

$$\bar{x}\Psi_{11\cdot 2}\bar{x}' = \bar{x}\Sigma_{11}^{-1}\bar{x}'. \tag{3.5}$$

If we now include the covariances, we obtain a loglikelihood function of a form similar to (3.2) in which V is defined by (2.1), $MW = N\bar{x}'\bar{x}$, and the term $M\log|\Sigma_{11}|$ does not appear. This model has been considered by a number of authors, e.g., Rao [42], Olkin and Shrikhande [40], and Gleser and Olkin [23].

In another context, suppose that a random sample of size N is observed from a p-variate normal distribution with covariance matrix Σ, and an additional sample of size M is observed on the first k (out of p) variates. Alternatively, this model can be viewed as a sample of size $N+M$ from a p-variate normal distribution, where the last $p-k$ (out of p) variates are missing from the last M observations. Let $(\bar{x}, \bar{y})$ denote the sample mean vector based on a sample of size N, where $\bar{x}$ refers to the mean vector of the first k variates. The sample mean vector on the first k variates from the additional sample of size M is denoted by $\bar{z}$.

The loglikelihood function that involves the means is proportional to

$$N(\bar{x}-\mu, \bar{y}-\nu)\Sigma^{-1}(\bar{x}-\mu, \bar{y}-\nu)' + M(\bar{z}-\mu)\Sigma_{11}^{-1}(\bar{z}-\mu)', \tag{3.6}$$

which is to be minimized with respect to μ and ν. Minimization of the first term of (3.6) with respect to ν is obtained directly from (3.3), which leads to the minimization problem

$$\min_{\mu}\left\{N(\bar{x}-\mu)\Sigma_{11}^{-1}(\bar{x}-\mu)' + M(\bar{z}-\mu)\Sigma_{11}^{-1}(\bar{z}-\mu)'\right\}.$$

This problem is straightforward and yields the minimum $(\bar{x}-\bar{z})\Sigma_{11}^{-1}(\bar{x}-\bar{z})'NM/(N+M)$.

If we now include the covariances, we obtain a loglikelihood function of the form (3.1), where V is defined by (2.1), and $W=\sum_1^M(z_\alpha-\bar{z})'(z_\alpha-\bar{z})/M+(\bar{x}-\bar{z})'(\bar{x}-\bar{z})N/(N+M)$. This problem was considered by Anderson [2], and also by Olkin and Sylvan [41] and by Giguère and Styan [21].

3.1. Differentiation

For simplicity of notation, write $A=(a_{ij})\equiv\Sigma_{11}$ and let A_{ij} denote the cofactor of a_{ij}, $i,j\leqslant k$. Using the differential forms in Section 2.1, we obtain

$$(2-\delta_{ij})\Big\{-N\frac{\Sigma_{ij}}{|\Sigma|}\,d\sigma_{ij}+N\big(\operatorname{tr}\Sigma^{-1}E_{ij}\Sigma^{-1}V\big)\,d\sigma_{ij}$$

$$-M\frac{A_{ij}}{|\Sigma_{11}|}\,da_{ij}\varepsilon_{ij}+\varepsilon_{ij}M\,da_{ij}\big(\operatorname{tr}\Sigma_{11}^{-1}E_{ij}\Sigma_{11}^{-1}W\big)\Big\}=0, \tag{3.7}$$

where $\varepsilon_{ij}=1$ if $i,j\leqslant k$, $\varepsilon_{ij}=0$, otherwise. Equation (3.7) can be written as the matrix equation

$$-N\Sigma^{-1}+N\Sigma^{-1}V\Sigma^{-1}-M\begin{pmatrix}\Sigma_{11}^{-1}&0\\0&0\end{pmatrix}+M\begin{pmatrix}\Sigma_{11}^{-1}W\Sigma_{11}^{-1}&0\\0&0\end{pmatrix}=0. \tag{3.8}$$

Pre- and postmultiplication by Σ in (3.8) yields

$$-N\Sigma+NV-M\Sigma\begin{pmatrix}\Sigma_{11}^{-1}&0\\0&0\end{pmatrix}\Sigma+M\Sigma\begin{pmatrix}\Sigma_{11}^{-1}W\Sigma_{11}^{-1}&0\\0&0\end{pmatrix}\Sigma=0,$$

which simplifies to the set of equations

$$\begin{aligned}-N\Sigma_{11}+NV_{11}-M\Sigma_{11}+MW&=0,\\-N\Sigma_{12}+NV_{12}-M\Sigma_{12}+MW\Sigma_{11}^{-1}\Sigma_{12}&=0,\\-N\Sigma_{22}+NV_{22}-M\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}+M\Sigma_{21}\Sigma_{11}^{-1}W\Sigma_{11}^{-1}\Sigma_{12}&=0.\end{aligned} \tag{3.9}$$

The equations (3.9) can be solved in sequence to yield

$$\hat{\Sigma}_{11} = \frac{1}{N+M}(NV_{11} + MW),$$

$$\hat{\Sigma}_{12} = \frac{1}{N+M}(NV_{11} + MW)V_{11}^{-1}V_{12},$$

$$\hat{\Sigma}_{22} = \frac{1}{N+M}\left[NV_{22} + MV_{22\cdot 1} + MV_{21}V_{11}^{-1}WV_{11}^{-1}V_{12}\right]. \tag{3.10}$$

The introductory comments of Section 2.1 apply here also, so that the solution $\hat{\Sigma}$ in (3.10) is a unique maximum.

3.2. *Transformations*

Since

$$\log|\Psi| = \log|\Psi_{11\cdot 2}| + \log|\Psi_{22}|$$

and

$$\begin{aligned}\operatorname{tr}\Psi V &= \operatorname{tr}\Psi_{11}V_{11} + 2\operatorname{tr}\Psi_{12}V_{21} + \operatorname{tr}\Psi_{22}V_{22} \\ &= \operatorname{tr}\Psi_{11\cdot 2}V_{11} + \operatorname{tr}\Psi_{22}V_{22} + \operatorname{tr}\Psi_{12}\Psi_{22}^{-1}\Psi_{21}V_{11} + 2\operatorname{tr}\Psi_{12}V_{21},\end{aligned}$$

we can rewrite (3.2) as

$$\begin{aligned}g(\Psi) = {} & N\log|\Psi_{22}| + (N+M)\log|\Psi_{11\cdot 2}| - N\operatorname{tr}\Psi_{11}V_{11} \\ & - N\operatorname{tr}\Psi_{12}V_{21} - N\operatorname{tr}\Psi_{21}V_{12} - N\operatorname{tr}\Psi_{22}V_{22} - M\operatorname{tr}\Psi_{11\cdot 2}W.\end{aligned} \tag{3.11}$$

Let

$$\Gamma = \Psi_{22}^{-1}\Psi_{21}, \qquad G = V_{21}V_{11}^{-1};$$

then (3.11) can be rewritten as

$$\begin{aligned}& N\log|\Psi_{22}| - N\operatorname{tr}\Psi_{22}V_{22\cdot 1} - N\operatorname{tr}\Psi_{22}(\Gamma+G)V_{11}(\Gamma+G)' \\ & \quad + (N+M)\log|\Psi_{11\cdot 2}| - \operatorname{tr}\Psi_{11\cdot 2}(NV_{11} + MW),\end{aligned} \tag{3.12}$$

where the maximum is over $\{\Psi : \Psi > 0\}$, or equivalently, $\{\Psi_{11\cdot 2}, \Psi_{22}, \Gamma : \Psi_{11\cdot 2} > 0,\ \Psi_{22} > 0\}$.

We now make use of the fact that if $Q_1 > 0$, $Q_2 > 0$, then

$$\operatorname{tr} Q_1 C Q_2 C' = \operatorname{tr} Q_1^{1/2} C Q_2 C' Q_1^{1/2} \geqslant 0, \tag{3.13}$$

with equality if and only if $C = 0$. It follows that $\operatorname{tr} \Psi_{22}(\Gamma + G)V_{11}(\Gamma + G)' \geqslant 0$ with equality at $\hat{\Gamma} = -G = -V_{21}V_{11}^{-1}$. The maximum of (3.12) with respect to Γ then leads to the sum of two terms like (2.3). The maximum over $\Psi_{22} > 0$ occurs at $\hat{\Psi}_{22} = V_{22\cdot 1}^{-1}$, and the maximum over $\Psi_{11\cdot 2} > 0$ occurs at $\hat{\Psi}_{11\cdot 2} = (N+M)(NV_{11} + MW)^{-1}$.

4. ESTIMATING THE MEANS WHEN THERE IS A RANK CONSTRAINT

In canonical form, let X and Y be $p \times N$ and $p \times M$ data matrices whose columns are independently distributed according to p-variate normal distributions with the same covariance matrix Σ and $EX = 0$, $EY = \Phi$, where Φ is of rank $r \leqslant p$ ($\leqslant M$). This model is considered by Anderson [1].

To obtain the maximum-likelihood estimators of Σ and Φ, we start with the likelihood function

$$c|\Sigma|^{-\frac{1}{2}(N+M)} \exp\left\{-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\left[XX' + (Y-\Phi)(Y-\Phi)'\right]\right\}, \tag{4.1}$$

where c is a normalizing constant. From Section 2, the maximum of (4.1) with respect to Σ (for fixed Φ) occurs at

$$\hat{\Sigma} = \frac{XX' + (Y-\Phi)(Y-\Phi)'}{N+M},$$

so that we need to determine

$$\begin{aligned} &\operatorname*{Min}_{\Phi} \left|XX' + (Y-\Phi)(Y-\Phi)'\right| \\ &\qquad = \operatorname*{Min}_{\Phi} |XX'| \left| I_p + (XX')^{-1/2}(Y-\Phi)(Y-\Phi)'(XX')^{-1/2} \right|. \end{aligned} \tag{4.2}$$

If we let

$$\tilde{Y} = (XX')^{-1/2}Y, \qquad \tilde{\Phi} = (XX')^{-1/2}\Phi,$$

then, except for the term $|XX'|$, (4.2) simplifies to

$$\operatorname*{Min}_{\tilde{\Phi}} \left| I_p + (\tilde{Y} - \tilde{\Phi})(\tilde{Y} - \tilde{\Phi})' \right| \tag{4.3}$$

over the region $\tilde{\Phi}: p \times M$ of rank r.

We wish to make use of a transformation for rectangular matrices that is equivalent to that of Fact 3.

FACT 3′. *If L is a $p \times M$ matrix of rank r, then there exists a $p \times r$ matrix T and an $M \times M$ orthogonal matrix $\Delta = (\Delta_1, \Delta_2)$ such that*

$$L = T\begin{pmatrix} I_r & 0 \end{pmatrix}\Delta' = T\Delta_1',$$

where $\Delta_1: M \times r$.

We apply this representation to

$$\tilde{\Phi} = T\Delta_1'$$

so that the determinant in (4.3) can be written as

$$\begin{aligned} \left| I_p + (\tilde{Y} - \tilde{\Phi})(\tilde{Y} - \tilde{\Phi})' \right| &= |I_p + \tilde{Y}\tilde{Y}' + TT' - T\Delta_1'\tilde{Y}' - \tilde{Y}\Delta_1 T'| \\ &= \left| \left(I_p + \tilde{Y}\tilde{Y}' - \tilde{Y}\Delta_1\Delta_1'\tilde{Y}' \right) + (T - \tilde{Y}\Delta_1)(T - \tilde{Y}\Delta_1)' \right| \\ &= \left| \left(I_p + \tilde{Y}\Delta_2\Delta_2'\tilde{Y}' \right) + (T - \tilde{Y}\Delta_1)(T - \tilde{Y}\Delta_1)' \right|. \end{aligned} \tag{4.4}$$

Since $I_p + \tilde{Y}\Delta_2\Delta_2'\tilde{Y}' > 0$, the minimum of (4.4) over T is achieved at $\hat{T} = \tilde{Y}\Delta_1$. This leads, with the use of the fact $|I_m + AB| = |I_n + BA|$ for $A: m \times n$, $B: n \times m$, to

$$\begin{aligned} \operatorname*{Min}_{\Delta_2} |I_p + \tilde{Y}\Delta_2\Delta_2'\tilde{Y}'| &= \operatorname*{Min}_{\Delta_2} |I_{M-r} + \Delta_2'\tilde{Y}'\tilde{Y}\Delta_2| \\ &= \operatorname*{Min}_{\Delta_2} \left| \Delta_2' \left(I_M + \tilde{Y}'\tilde{Y} \right) \Delta_2 \right|. \end{aligned} \tag{4.5}$$

To obtain (4.5) we require the following.

FACT 7. *If A is an $m \times m$ positive semidefinite matrix with ordered characteristic roots $\alpha_1 \geq \cdots \geq \alpha_m$, and $\mathscr{U}$ is the set of $k \times m$ $(k \leq m)$ matrices U satisfying $UU' = I_k$, then*

$$\operatorname*{Min}_{U \in \mathscr{U}} |UAU'| = \prod_1^k \alpha_{m-i+1}.$$

(See, e.g., [11, Theorem 10, p. 132].)

An application of Fact 7 to (4.5) yields a minimum of $\prod_{r+1}^M \lambda_i$, where $\lambda_1 \geq \cdots \geq \lambda_M > 1$ are the ordered characteristics root of $I_M + \tilde{Y}'\tilde{Y} = I_M + Y'(XX')^{-1}Y$, or equivalently, the roots of $|XX' + YY' - \lambda XX'| = 0$. This minimum is achieved by $\hat{\Delta}_2 = (c'_{r+1}, \ldots, c'_M)$, where c_i is the vector satisfying $[I_M + Y'(XX')^{-1}Y]c'_i = \lambda_i c'_i$ and $c_i c'_i = 1$, $i = r+1, \ldots, M$.

Recall from Fact 3′ and the minimization of (4.4) that $\tilde{\Phi} = T\Delta'_1$ and that $\hat{T} = \tilde{Y}\Delta_1$. The matrix Δ_1 still needs to be determined. However, since $\Delta = (\Delta_1, \Delta_2)$ is orthogonal, and $\hat{\Delta}_2$ satisfying $\hat{\Delta}'_2\hat{\Delta}_2 = I$ is determined, any matrix $\hat{\Delta}_1$ satisfying $\hat{\Delta}'_1\hat{\Delta}_1 = I_r$, $\hat{\Delta}'_1\hat{\Delta}_2 = 0$ yields a solution.

In summary, the maximized likelihood is

$$c\left\{\frac{e^{-p}(N+M)^p}{|XX'|\prod_{r+1}^M \lambda_i}\right\}^{(N+M)/2} \tag{4.6}$$

The problem of this section is considered by Healy [25]. His procedure is to first make a series of transformations motivated by the rank condition on Φ. The effect is to transform the model to a canonical form, upon which he performs the maximizations. However, by first maximizing with respect to the covariances (as in the above derivation), the required transformations (and proof) become considerably simpler. The model (4.1), without the rank restrictions, is discussed by Calvert and Seber [15].

5. ESTIMATING THE MEANS IN A REGRESSION CONTEXT

In this model the random $p \times M$ matrix Y has a mean BZ, where the known matrix $Z: k \times M$ is of rank k, and $B: p \times k$ is a matrix of parameters. The columns of Y are independent with common covariance matrix Σ. The loglikelihood function is proportional to the negative of

$$M \log|\Sigma| + \operatorname{tr} \Sigma^{-1}(Y - BZ)(Y - BZ)', \tag{5.1}$$

which is to be minimized with respect to B. Write

$$Y - BZ = \left[Y - YZ'(ZZ')^{-1}Z\right] + \left[YZ'(ZZ')^{-1}Z - BZ\right] \equiv (A_1 + A_2),$$

and note that $A_1A_2' = 0$. Consequently,

$$\operatorname{tr}\Sigma^{-1}(Y - BZ)(Y - BZ)' = \operatorname{tr}\Sigma^{-1}A_1A_1' + \operatorname{tr}\Sigma^{-1}A_2A_2'. \tag{5.2}$$

The first term

$$\operatorname{tr}\Sigma^{-1}A_1A_1' = \operatorname{tr}\Sigma^{-1}Y\left[I - Z'(ZZ')^{-1}Z\right]Y'$$

is independent of B. The second term becomes

$$\operatorname{tr}\Sigma^{-1}A_2A_2' = \operatorname{tr}\Sigma^{-1}\left[YZ'(ZZ')^{-1} - B\right]ZZ'\left[YZ'(ZZ')^{-1} - B\right]' \tag{5.3}$$

which, using (3.13), is minimized at

$$\hat{B} = YZ'(ZZ')^{-1}.$$

6. ESTIMATING MEANS UNDER LINEAR CONSTRAINTS

Suppose Y is a $p \times M$ ($p \leqslant M$) random matrix with mean

$$EY = Z_1BZ_2, \tag{6.1}$$

where B is an unknown $q \times m$ matrix $Z_1: p \times q$ is of rank q, $Z_2: m \times M$ is of rank m, and Z_1 and Z_2 are known. The columns of Y are independent with common covariance matrix Σ.

The problem is to find the maximum-likelihood estimate of B subject to the constraint

$$Z_3BZ_4 = B_0, \qquad \text{known}, \tag{6.2}$$

where $Z_3: v \times q$ is of rank v, $Z_4: m \times u$ is of rank u, and Z_3 and Z_4 are known.

Variants of this model have a long history. It was considered in its present general form by Gleser and Olkin [23].

We show how, by a series of transformations, this model can be reduced to a simple canonical form.

The main transformation is that of Fact 3′ as applied to different matrices. First let

$$Z_1' = T_1'\begin{pmatrix} I_q & 0 \end{pmatrix}\Gamma_1', \qquad Z_2 = T_2\begin{pmatrix} I_m & 0 \end{pmatrix}\Gamma_2, \tag{6.3}$$

where $\Gamma_1: p \times p$ and $\Gamma_2: m \times m$ are orthogonal and T_1 and T_2 are nonsingular. Then (6.1) becomes

$$EY = \Gamma_1\begin{pmatrix} I_q \\ 0 \end{pmatrix}T_1 B T_2\begin{pmatrix} I_m & 0 \end{pmatrix}\Gamma_2. \tag{6.4}$$

Write

$$B^* = T_1 B T_2, \qquad Y^* = \Gamma_1' Y \Gamma_2'.$$

Since T_1 and T_2 are known and nonsingular, minimization with respect to B^* is equivalent to minimization with respect to B. Then (6.4) becomes

$$EY^* = E\left(Y_1^*, Y_2^*\right) = \begin{pmatrix} I_q \\ 0 \end{pmatrix}B^*(I_m, 0) = \begin{pmatrix} B^* & 0 \\ 0 & 0 \end{pmatrix}, \tag{6.5}$$

where $Y_1^*: p \times m$ and $Y_2^*: p \times (M - m)$.

Now consider the constraint equation (6.2) and let

$$Z_3 T_1^{-1} = T_3\begin{pmatrix} I_v & 0 \end{pmatrix}\Gamma_3, \qquad T_2^{-1} Z_4 = \Gamma_4\begin{pmatrix} I_u \\ 0 \end{pmatrix}T_4,$$

where $\Gamma_3: q \times q$ and $\Gamma_4: M \times M$ are orthogonal, and T_3 and T_4 are nonsingular. Then (6.2) becomes

$$Z_3 B Z_4 = \left(Z_3 T_1^{-1}\right)\left(T_1 B T_2\right)\left(T_2^{-1} Z_4\right) = T_3\begin{pmatrix} I_v & 0 \end{pmatrix}\Gamma_3 B^* \Gamma_4\begin{pmatrix} I_u \\ 0 \end{pmatrix}T_4 = B_0.$$

Since T_3 and T_4 are known, the constraint equation is equivalent to

$$\begin{pmatrix} I_v & 0 \end{pmatrix}\left(\Gamma_3 B^* \Gamma_4\right)\begin{pmatrix} I_u \\ 0 \end{pmatrix} = T_3^{-1} B_0 T_4^{-1} \equiv B_0^*. \tag{6.6}$$

A final simplification is achieved by letting

$$\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2) = \begin{pmatrix} \Gamma_3 & 0 \\ 0 & I_{p-q} \end{pmatrix} Y^* \begin{pmatrix} \Gamma_4 & 0 \\ 0 & I_{M-m} \end{pmatrix}.$$

These transformations now yield the following model: $\tilde{Y}_1$ is a random $p \times m$ matrix; $\tilde{Y}_2$ is a random $p \times (M-m)$ matrix with means

$$E\tilde{Y}_1 = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \\ 0 & 0 \end{pmatrix}, \qquad E\tilde{Y}_2 = 0.$$

The matrices $\tilde{Y}_1$ and $\tilde{Y}_2$ are independently distributed; the columns of $\tilde{Y}$ have a common covariance matrix $\Psi = \Gamma_1 \Sigma \Gamma_1'$. The constraint (6.6) now becomes $\theta_{11} = \theta_{11}^0$, which can be taken to be 0 without loss of generality.

The likelihood function when θ_{12} is arbitrary can be obtained from Section 5, say. The likelihood function when $\theta_{11} = 0$ reduces to the model of Section 3.

7. ESTIMATING VARIANCES WHEN CORRELATIONS ARE FIXED

The starting point in this model is $f(\Sigma; V)$ of Section 2. The problem is to estimate the variances $\sigma_{11}, \ldots, \sigma_{pp}$ when the correlations ρ_{ij} are fixed. Using the transformation of Fact 4, let

$$\Sigma = D_\sigma P D_\sigma, \tag{7.1}$$

where $D_\sigma = D(\sigma_1, \ldots, \sigma_p)$, $\sigma_i^2 = \sigma_{ii}$, $i = 1, \ldots, p$, and $P = (\rho_{ij})$ is a correlation matrix. Similarly, let

$$V = D_v R D_v, \tag{7.2}$$

where $D_v = D(v_1, \ldots, v_p)$, $v_i^2 = v_{ii}$, $i = 1, \ldots, p$, and $R = (r_{ij})$ is the sample correlation matrix. Then the likelihood (2.2) (except for constants) is

$$f(\Sigma; V) = -\log|D_\sigma|^2 - \log|P| - \operatorname{tr} D_\sigma^{-1} P_\sigma^{-1} D_\sigma^{-1} D_v R D_v. \tag{7.3}$$

Let

$$\tau_i = v_i/\sigma_i, \quad i = 1, \ldots, p, \qquad D_\tau = D_v D_\sigma^{-1}.$$

Then (7.3) becomes

$$f(\Sigma; V) = 2\sum \log \tau_i - \operatorname{tr} D_\tau P^{-1} D_\tau R + (-\log|P| - 2\log|D_v|). \quad (7.4)$$

The last term is independent of τ and can be ignored in the maximization. The second term is

$$\operatorname{tr} D_\tau P^{-1} D_\tau R = \sum \tau_i \rho^{ij} \tau_j r_{ji} = (\tau_1, \ldots, \tau_p) A (\tau_1, \ldots, \tau_p)',$$

where $A = (a_{ij})$, $a_{ij} = \rho^{ij} r_{ij}$, $P^{-1} = (\rho^{ij})$. Since P and R are positive definite, the matrix A is positive definite (see e.g. [11, p. 95]). Consequently, the maximum-likelihood estimator of τ is obtained by maximizing

$$2\sum_i \log \tau_i - (\tau_1, \ldots, \tau_p) A (\tau_1, \ldots, \tau_p)'$$

over the region $\tau_i > 0$, $i = 1, \ldots, p$. Setting the derivatives with respect to $\tau_1, \ldots, \tau_p$ equal to zero yields the matrix equation [49]

$$\left(\frac{1}{\tau_1}, \ldots, \frac{1}{\tau_p} \right) = (\tau_1, \ldots, \tau_p) A. \quad (7.5)$$

This equation has an interesting history in another context. If $e = (1, \ldots, 1)$, then (7.5) can be written as

$$e D_\tau^{-1} = e D_\tau A, \quad (7.6)$$

from which we obtain

$$e = e(D_\tau A D_\tau). \quad (7.7)$$

Equation (7.7) has the following interpretation: Given a positive definite matrix A, find a scaling matrix D_τ of positive elements such that $D_\tau A D_\tau$ is doubly stochastic. This problem originally arose for the case when A has nonnegative elements. Many variants and extensions have been noted. The problem has a long history and has received considerable attention in the numerical-analysis literature. For a review of results see [30]. Of importance in (7.5) is the fact that if A is positive definite, then there exists a positive diagonal matrix D_τ satisfying (7.7). Numerically this can be obtained as a

solution of the extremal problem

$$\operatorname*{Max}_{\prod x_i = 1} xAx'.$$

It can also be obtained iteratively by alternately normalizing rows and columns: $A_k = D_{k1}A_{k-1}D_{k2}$, $A_0 \equiv A$, $k = 1,2,\ldots$.

Estimation of the variances when the correlations are fixed, and when the correlation matrix has a linear structure, was studied by Styan [48].

8. A PROBLEM ON IDEMPOTENCY

Consider the regression model $y = \beta X + \varepsilon$, where $\beta: 1 \times p$ is unknown, $X: p \times n$ is a known matrix of rank p, and ε has an n-variate normal distribution with mean vector zero and covariance matrix $\sigma^2 I_n$. Theil and Schweitzer [53] obtain a quadratic estimate yAy' of σ^2 that is nonnegative, has a distribution independent of β, and minimizes $E(yAy' - \sigma^2)^2$. This problem is equivalent to the extremal problem

$$\operatorname*{Min}_{\mathscr{A}} \left[2\operatorname{tr}A^2 + (1 - \operatorname{tr}A)^2\right] \tag{8.1}$$

over the region $\mathscr{A} = \{A: A \geqslant 0,\ AX' = 0,\ A: n \times n,\ X: p \times n \text{ of rank } p\}$. Theil and Schweitzer [53] resolve the minimization by using Lagrangian multipliers for the constraint equation $AX = 0$. Calvert and Seber [15] embed the problem in a more general extremal setting, which they solve. Neudecker [35] offers another solution. We show how to reduce the problem by a series of transformations that has the effect of removing the constraint, after which the solution is readily obtained.

Using Fact 3′, let $X = T(I_p, 0)K$, where $T: p \times p$ is nonsingular and K is an $n \times n$ orthogonal matrix. Write $\tilde{A} = KAK'$ partitioned as $\tilde{A} = (\tilde{A}_{ij})$, $i, j = 1,2$, $\tilde{A}_{11}: p \times p$, $\tilde{A}_{22}: (n-p) \times (n-p)$. The condition $AX' = 0$ becomes

$$\tilde{A}\begin{pmatrix} I_p \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{A}_{11} \\ \tilde{A}_{21} \end{pmatrix} = 0, \tag{8.2}$$

so that

$$\operatorname{tr}A^2 = \operatorname{tr}\tilde{A}_{22}^2, \qquad \operatorname{tr}A = \operatorname{tr}\tilde{A}_{22}.$$

Then (8.1) reduces to

$$\min_{\tilde{A}_{22} \geq 0} \left\{ 2 \operatorname{tr} \tilde{A}_{22}^2 + \left(1 - \operatorname{tr} \tilde{A}_{22}\right)^2 \right\}, \tag{8.3}$$

which, from Fact 5, is equivalent to

$$\min_{\theta_i \geq 0} \left\{ 2 \sum_i \theta_i^2 + \left(1 - \sum_i \theta_i\right)^2 \right\}, \tag{8.4}$$

where $\theta_1, \ldots, \theta_{n-p}$ are the characteristic roots of $\tilde{A}_{22}$. The minimum of (8.4) is $2/(n-p+2)$ and is achieved by $\hat{\theta}_1 = \cdots = \hat{\theta}_{n-p} = 1/(n-p+2)$.

REFERENCES

1 T. W. Anderson, Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Ann. Math. Statist.* 22:327–351 (1951).

2 T. W. Anderson, Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, *J. Amer. Statist. Assoc.* 52:200–203 (1957).

3 T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958; 2nd ed., 1984.

4 T. W. Anderson, Statistical inference for covariance matrices with linear structure, in *Multivariate Analysis II*, (P. R. Krishnaiah, Ed.), Academic, New York, 1969, pp. 55–66.

5 T. W. Anderson, Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices, in *Essays in Probability and Statistics* (R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, and K. J. C. Smith, Eds.), Univ. of North Carolina Press, Chapel Hill, N.C., 1970, pp. 1–24.

6 S. Andersson, Invariant normal models, *Ann. Statist.* 3:132–154 (1975).

7 S. F. Arnold, Application of the theory of products to problems of certain patterned covariance matrices, *Ann. Statist.* 1:682–699 (1973).

8 S. F. Arnold, Applications of products to the generalized compound symmetry problem, *Ann. Statist.* 4:227–233 (1976).

9 O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, Wiley, New York, 1978.

10 M. S. Bartlett, On the theory of statistical regression, *Proc. Roy. Soc. Edinburgh* 53:260–283 (1933).

11 R. Bellman, *Introduction to Matrix Analysis*, 2nd ed., McGraw-Hill, New York, 1970.

12 A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, Wiley-Interscience, New York, 1974; reprint ed. with corrections, Robert E. Krieger, Huntington, N.Y., 1980.
13 E. Bodewig, *Matrix Calculus*, 2nd ed., North-Holland, Amsterdam, 1959.
14 T. L. Boullion and P. L. Odell, *Generalized Inverse Matrices*, Wiley, New York, 1971.
15 B. Calvert and G. A. F. Seber, Minimization of functions of a positive semidefinite matrix A subject to $AX = 0$, *J. Multivariate Anal.* 8:274–281 (1978).
16 W. J. Deemer and I. Olkin, The Jacobians of certain matrix transformations useful in multivariate analysis, *Biometrika* 38:345–367 (1951).
17 P. S. Dwyer, Some applications of matrix derivatives in multivariate analysis, *J. Amer. Statist. Assoc.* 62:607–625 (1967).
18 P. S. Dwyer and M. S. Macphail, Symbolic matrix derivatives, *Ann. Math. Statist.* 19:517–534 (1948).
19 M. Erlandsen, Affine hypotheses in the mean and in the covariance in multivariate normal families, *Scand. J. Statist.* 8:10–16 (1981).
20 R. A. Frazer, W. J. Duncan and A. R. Collar, *Elementary Matrices and Some Applications to Dynamics and Differential Equations*, University Press, Cambridge, England, 1947.
21 M. A. Giguère and G. P. H. Styan, Multivariate normal estimation with missing data on several variates in, *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the Eighth European Meeting of Statisticians* (Prague, August 1974), D. Reidel, Dordrecht, Netherlands, 1978, Vol. B, pp. 129–139.
22 N. C. Giri, *Multivariate Statistical Inference*, Academic, New York, 1977.
23 L. J. Gleser and I. Olkin, Linear models in multivariate analysis, in *Essays in Probability and Statistics* (R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, and K. J. C. Smith, Eds.), Univ. of North Carolina Press, Chapel Hill, N.C., 1969, pp. 267–292.
24 A. Graham, *Kronecker Products and Matrix Calculus: With Applications*, Halsted, New York, 1981.
25 J. D. Healy, Maximum likelihood estimation of a multivariate linear functional relationship, *J. Multivariate Anal.* 10:243–251 (1980).
26 C. G. Khatri, Letter to the editor, *Amer. Statist.* 33:92 (1979).
27 A. M. Kshirsagar, *Multivariate Analysis*, Marcel Dekker, New York, 1972.
28 P. C. Mahalanobis, R. C. Bose and S. N. Roy, Normalisation of statistical variates and the use of rectangular coordinates in the theory of sampling distributions, *Sankhyā* 3:1–40 (1937).
29 K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic, London, 1979.
30 A. W. Marshall and I. Olkin, Scaling of matrices to achieve specified row and column sums, *Numer. Math.* 12:83–90 (1968).
31 R. P. McDonald and H. Swaminathan, A simple matrix calculus with applications to multivariate analysis, *General Systems* 18:37–54 (1973).
32 R. J. Muirhead, *Aspects of Multivariate Statistical Analysis*, Wiley, New York, 1982.

33 D. G. Nel, On matrix differentiation in statistics, *South African Statist. J.* 14:137–193 (1980).

34 H. Neudecker, Some theorems on matrix differentiation with special reference to Kronecker matrix products, *J. Amer. Statist. Assoc.* 64:953–963 (1969).

35 H. Neudecker, A comment on "Minimization of functions of a positive semidefinite matrix A subject to $AX = 0$," *J. Multivariate Anal.* 10:135–139 (1980).

36 J. von Neumann, Some matrix inequalities and metrization of matrix space, *Tomsk Univ. Rev.* 1:286–300 (1937); reprinted in *John von Neumann: Collected Works*, (A. H. Taub, Ed.), Pergamon, New York, 1962, Vol. 4, pp. 205–218.

37 I. Olkin, Testing and estimation for structures which are circularly symmetric in blocks, in *Multivariate Statistical Inference* (D. G. Kabe and R. P. Gupta, Eds.), North-Holland, Amsterdam, 1973, pp. 183–195.

38 I. Olkin and S. J. Press, Testing and estimation for a circular stationary model, *Ann. Math. Statist.* 40:1358–1373 (1969).

39 I. Olkin and A. R. Sampson, Jacobians of matrix transformations and induced functional equations, *Linear Algebra Appl.* 5:257–276 (1972).

40 I. Olkin and S. S. Shrikhande, On a modified T^2 problem, *Ann. Math. Statist.* 25:80 (1954).

41 I. Olkin and M. Sylvan, Correlational analysis when some variances and covariances are known, in *Multivariate Analysis — IV* (P. R. Krishnaiah, Ed.), North-Holland, Amsterdam, 1977, pp. 175–191.

42 C. R. Rao, On some problems arising out of discrimination with multiple characters, *Sankhyā* 9:343–366 (1949).

43 C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York, 1965; 2nd ed., 1973.

44 C. R. Rao and S. K. Mitra, *Generalized Inverses of Matrices and Its Applications*, Wiley, New York, 1971.

45 G. S. Rogers, *Matrix Derivatives*, Marcel Dekker, New York, 1980.

46 E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener, *Math. Ann.* 63:433–476 (1907).

47 D. W. Smith, A simplified approach to the maximum likelihood estimation of the covariance matrix, *Amer. Statist.* 32:28–29 (1978).

48 G. P. H. Styan, Multivariate normal inference with correlation structure, Ph.D. Dissertation, Dept. of Mathematical Statistics, Columbia Univ., New York, 1969.

49 G. P. H. Styan, Hadamard products and multivariate statistical analysis, *Linear Algebra Appl.* 6:217–240 (1973).

50 T. H. Szatrowski, Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances, *Ann. Statist.* 8:802–810 (1980).

51 T. H. Szatrowski, Relative efficiencies of estimates using patterned covariance or correlations in the multivariate normal estimation problem, *Ann. Inst. Statist. Math.* 34:299–307 (1982).

52 A. C. Tamhane, Letter to the Editor, *Amer. Statist.* 33:92–93 (1979).

53 H. Theil and A. Schweitzer, The best quadratic estimator of the residual variance in regression analysis, *Statist. Neerlandica* 15:19–23 (1961).

54 C. M. Theobald, An inequality with application to multivariate analysis, *Biometrika* 62:461–466 (1975).
55 O. Toeplitz, Die Jacobische Transformation der quadratischen Formen von unendlich vielen Veränderlichen, *Nachr. Kgl. Gesellschaft Wiss. Göttingen Math.-Phys. Kl.*, 1907, pp. 101–109.
56 D. S. Tracy and R. P. Singh, Some applications of matrix differentiation in the general analysis of covariance structures, *Sankhyā Ser. A* 37:269–280 (1975).
57 D. F. Votaw, Jr., Testing compound symmetry in a normal multivariate distribution, *Ann. Math. Statist.* 19:447–473 (1948).
58 G. S. Watson, A note on maximum likelihood, *Sankhyā* 26:303–304 (1964).
59 S. S. Wilks, Certain generalizations in the analysis of variance, *Biometrika* 24:471–494 (1932).

Received 29 May 1984; revised 15 February 1985